MULTIMEDIA UNIVERSITY®

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2022/2023

HPB2031 –BIOINFORMATICS ALGORITHMS I

(All sections / Groups)

4 JULY 2023
9:00 AM- 11:00 AM
(2 hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 4 pages, including this cover page.
2. You are required to attempt all questions. All questions carry equal marks (10).
3. Write all your answers in the Answer Booklet provided.
4. You may use a calculator.

Question 1

a) Given the alignment matrix below:

		G	C	A	T	G	C
	0	-3	-6	-9	-12	-15	-18
G	-3	3	0	-3	-6	-9	-12
A	-6	0	1	3	0	-3	-6
T	-9	-3	-1	0	6	3	0
T	-12	-6	-4	-3	3	4	2
A	-15	-9	-7	-1	0	2	2
C	-18	-12	-6	-4	-2	-1	5

Determine:

i. whether the matrix is either global or local alignment. [1 mark]

ii. Scoring for mismatches (transition and transversion), match and gap. [2 marks]

iii. Optimal alignment [1 mark]

b) A research student searched for intra-species (same species) *16S* rDNA homologs of *Escherichia coli* in NCBI by using *blastn* program against nr database. Although the *16S* sequences of *E. coli* are abundantly available in the public database, the student could not get the expected hits. Propose the possible reason and suggest a solution. [2 marks]

c) Why are PAM and BLOSUM only represented at either the upper or lower triangular matrices but not both? [1 mark]

d) How is FASTA alignment algorithm different than BLAST algorithm? [1 mark]

e) Given two orthologous sequences, each retrieved from yeast and human, which substitution matrices (BLOSUM 90 and PAM 1) will be suitable to be used for an alignment? Justify your answer. [2 marks]

Continued .....

Question 2

a) Why is Illumina sequencing known as the only high throughput sequencing platform? What is the technology used for the purpose? [2 marks]

b) How many kmer exist in the sequence of "AATGGTACAAT" with the kmer size of 3? Illustrate all the kmer. [2 marks]

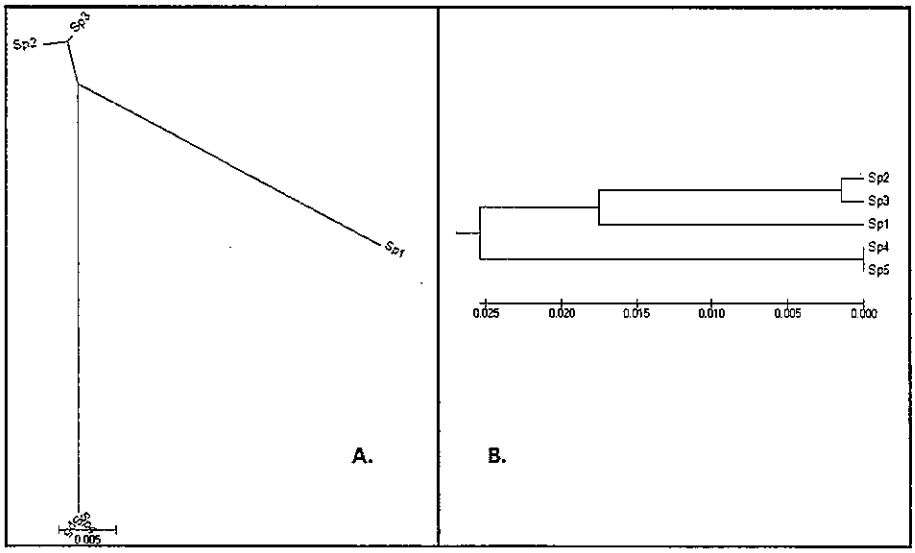
c) A kmer size of 16 was set to generate a genome assembly. However, the assembly was found to be made up of many chimera (eg: genomic fusions, abundant number of contigs, no BLAST hit)? In your opinion, what is the contributing factor for the observation? Suggest a solution. [2 marks]

d) How do PacBio-generated data and Illumina-yielded data complement each other in genome assembly? [2 marks]

e) A scientist achieved highly fragmented draft genome by using **Canu** assembler to assemble **Illumina** generated data. What is the possible contributing factor for the observation? Suggest a solution to improve the assembly. [2 marks]

Question 3

a)



The figure above shows the phylogenetic trees of five species reconstructed using different approaches. Describe and explain each of the approaches used in A and B. What is the rooting method used in B? [3 marks]

Continued .....

b) Given the distance matrix below (N = 6), show the Neighbor-Joining steps to derive the corresponding phylogenetic tree. Calculate

i. r(i) [1.5 marks]
ii. then new matrix M(i,j) [2.5 marks]
iii. choose neighbors as new 'internal node' [1 mark]
iv. calculate distance matrix with new node [2 marks]

You are expected to complete only ONE iteration.

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Question 4

a) List the two types of interactions between proteins. [1 mark]

b) How is phylogenetic profile used to infer interaction between proteins? Why is phylogenetic profile different than observing the co-expression of genes? [2 marks]

c) Name and explain any two prediction channels of protein-protein interaction in STRING. [2 marks]

d) What are the three tiers in PGDB? Describe briefly. [3 marks]

e) What is the purpose of performing pathway enrichment analysis? List ONE statistical method. [2 marks]

END OF PAPER